Improving Classification Accuracy of Decision Trees for Different Abstraction Levels of Data

Mina Jeong, Mokpo National University, Korea
Doheon Lee, Korea Advanced Institute of Science and Technology, Korea

ABSTRACT

Classification is an important problem in data mining. Given a database of records, each tagged with a class label, a classifier generates a concise and meaningful description for each class that can be used to classify subsequent records. Since the data is collected from disparate sources in many actual data mining environments, it is common to have data values in different abstraction levels. This article introduces the multiple abstraction level problem in decision tree classification, and proposes a method to deal with it. The proposed method adopts the notion of fuzzy relation for solving the multiple abstraction level problem. The experimental results show that the proposed method reduces classification error rates significantly when multiple abstraction levels of data are involved.

Keywords: abstraction levels; classification; data mining; data quality; decision tree

INTRODUCTION

Classification is one of the most widely used tasks in data mining. It consists of the construction phase and the assignment phase. The construction phase builds a classification model based on a given training set. The assignment phase utilizes the classification model to assign class labels to target records. Among many classification models proposed in the literature, decision trees are especially attractive in data mining environments due to their intuitive representation, which is easy for humans to understand (Gehrke et al., 2000; Gehrke et al., 1999; Berry & Linoff, 1997; Quinlan, 1993; Mehta et al., 1996; Shafer et al., 1996).

The training set is commonly collected from disparate sources in actual data mining environments. Information gathering agents access local sources, extract relevant data items, and deliver them to the central repository. Since the information gathering agents do not have any prior
knowledge about global data distribution, the collected data in the training set is apt to be expressed in different abstraction levels. For example, a sales item description is expressed as “Coke” in a local table, while it is expressed as “Diet Coke 1.5 P.E.T.” in another table. We call this the multiple abstraction level problem. Uncontrolled data entry by human users can also cause the same problem. Though sophisticated user interfaces of data entry software and employee training programs can help to standardize data entry, the problem cannot be eliminated due to mistakes and/or unavailable information (English, 1999).

The multiple abstraction levels can cause a severe problem in decision tree classification. Suppose that we have built a decision tree from multiple abstraction levels of training data as shown in Figure 1. Notice that the abstraction level of “Far-East” and “Mid-East,” and that of “East” and “West” are different. Actually, “Far-East” and “Mid-East” are specializations of “East.” Also suppose that a class label is assigned to a target record such that (Mid-East, 85k, Male). One choice is to follow the third branch of the root node, and assigns “E2” to the target record. However, since “Mid-East” belongs to “East,” the other choice is to follow the first branch of the root node and in turn, the second branch of the “Income” node and assign “E1” to the target record. As a result, the decision tree yields two contradictory assignments.

As this sort of data quality problem has been a subject of long-standing discussions (English, 1999; Wang et al., 1995), data cleansing tools have been developed to improve the data quality on the market (Trilium Software systems, 1998; Vality Technology Inc., 1998; Williams, 1997). At first glance, it seems that there are two options to remedy the problem by using such tools. One option is to equalize abstraction levels by generalizing data values that are too specific. It is obvious that more specific data values yield more informative decision trees (Quinlan, 1993). Thus, this option results in the loss of useful information. The other option might be equalizing abstraction levels by specializing data values that are too general. Since it requires extracting additional information from the sources, it is hard or even impossible in actual data mining environments. Thus, the second option is also inapplicable due to the lack of information. Furthermore, it is also hard to determine how many levels we should consider generalizing or specializing data values in both options. Consequently, the existing data cleansing tools cannot provide satisfactory solutions for the problem.

Figure 1. An example decision tree with multiple abstraction levels of data
Data Field for Hierarchical Clustering
Shuliang Wang, Wenyan Gan, Deyi Li and Deren Li (2013). *Developments in Data Extraction, Management, and Analysis* (pp. 303-324).
www.igi-global.com/chapter/data-field-hierarchical-clustering/70803?camid=4v1a

Constrained Cube Lattices for Multidimensional Database Mining
Alain Casali, Sébastien Nedjar, Rosine Cicchetti and Lotfi Lakhal (2010). *International Journal of Data Warehousing and Mining* (pp. 43-72).
www.igi-global.com/article/constrained-cube-lattices-multidimensional-database/44958?camid=4v1a